Predicting Cancer Heterogeneity from One-shot Biopsy

Kazuki Ikeda*, Teppei Nakano
*Department of Physics, Osaka University, Toyonaka, Osaka 560-0043, Japan
kikeda@het.phys.sci.osaka-u.ac.jp
Keio University School of Medicine, Tokyo 160-8582, Japan
keiohigh2nd@gmail.com

Abstracts

Cancer heterogeneity leads to a wide variety of responses to therapy. However this fact is not mainly considered in clinical situations. With the advance of genome sequence technology as represented by single-cell RNA sequencing, we need to take heterogeneity concepts into clinical treatments and examinations. Here we focus on a biopsy from cancer. For instance, lung biopsy is important for molecular-target drugs because the drug only cure molecular positive tumors such as EGFR. EGFR-targeted drug can cure EGFR positive adenocarcinoma with the success rate 80% but the rest of 20% EGFR positive adenocarcinoma are not effective. This is due to existing biopsy cannot extract EGFR gene information correctly because of heterogeneity. Using 430 single-cell RNA sequencing data, we prove a random biopsy cannot verify the expression of EGFR with Kolmogorov-Smirnov test (p-value 0.64±0.02) from real patients' data (the largest test ever experimented). In second experiment, we build heterogeneity prediction model from one time biopsy since we want to predict heterogeneity with the fewest number of biopsies. Our model succeeded in predicting heterogeneity of 3 genes, which suggests cancer heterogeneity is predictable from one shot biopsy in specific condition.

1. Introduction

Heterogeneity is the differences between tumors of the same type in different patients, and between cancer cells within a tumor. Detection of minor, genetically distinct subpopulations within tumors is a key challenge in cancer genomics [1,2] because the subpopulations interacts each other to survive from anti-cancer drugs and severe microenvironments. However we are far from understanding the dynamics of the interactions among subpopulations. Thus at the moment, we cannot take cancer heterogeneity into clinical situations.

Surprisingly as for prostate cancer, Gleason grading system has been developed from the 1960s [3] and it is still widely used for the prognosis of cancer [4]. Gleason grading is based on pathological images to examine heterogeneity. It collects at least two samples from a prostate biopsy and the score is calculated from histological patterns from two tissue slides. Considering

cancer heterogeneity into clinical staging, Gleason score is one of the most useful benchmark in cancer. However other cancers such as lung cancer does not have grading system considering heterogeneity because lung adenocarcinoma is similar morphology in a tumor.

The aim of this work is to extract cancer heterogeneity feature from one time biopsy using single-cell RNA sequencing (Single-cell RNA-seq) not histology. In addition to the morphology, prostate cancer needs biopsy more than two times to extract heterogeneity. However biopsy from lung should be done at one time. This is due to lung biopsy is more dangerous than from prostate biopsy since it can cause pneumothorax and pneumonia [5]. So in terms of biopsy risks, it is better to practice lung biopsy at one time to get heterogeneity features.

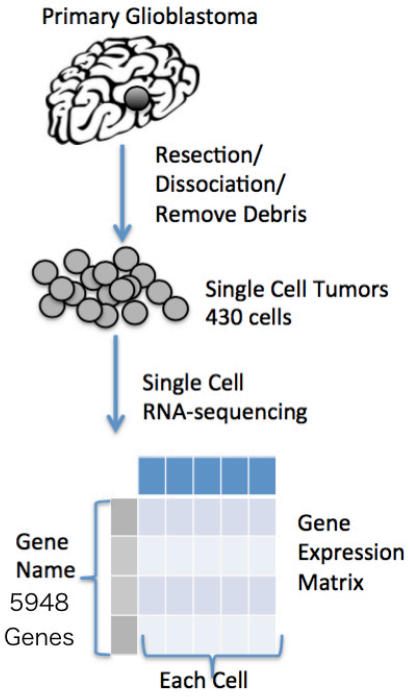

Fig 1. Datasets we analyzed

The 1st experiment of this paper focuses on epidermal growth factor receptor (EGFR) genes. In clinical situation, EGFR is important because EGFR-targeted drugs do cure EGFR positive adenocarcinoma with 80% response rate [6,7]. However the rest of 20% EGFR positive is not cured because of the heterogeneity. Therefore the main challenge we are interested in is how to diagnose and treat such ineffective lung cancer from one-shot biopsy. Genome sequencing costs money and time so we need to examine cancer heterogeneity with the fewest number of biopsies.

Currently single-cell RNA-seq made possible to reveal gene expression of each cancer cell. We used the data to 1) prove that existing biopsy is unstable to verify EGFR positive and 2) make a model that predicts the variance of gene expression from other genes.

2. Methods & Experiment

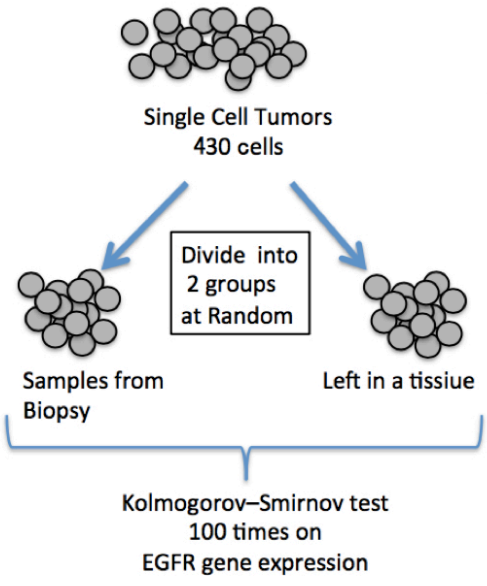

Fig 2.a Experiment 1

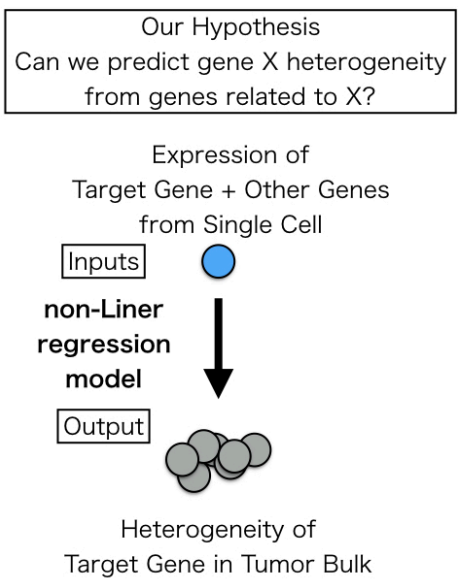

Fig 2.b Experiment 2

Single-cell RNA-seq technology has a big effect on caner genome study [8]. We used gene expression data from the research of glioblastoma heterogeneity (Fig 1.b) [9]. SMART-seq protocol was implemented to generate single cell full length transcriptomes and sequenced using 25 bp paired end reads. Cells were also cultured in serum free conditions to generate gliomasphere cell lines which were then differentiated using 10% serum (DGC). Population RNA-seq was performed on these sample. The initial dataset included 875 RNA-seq libraries (576 single glioblastoma cells, 192 single gliomasphere cells, 5 tumor population controls, 6 population libraries from GSC and DGC samples). Data was processed as described below using RSEM for quantification of gene expression. 5,948 genes with the highest composite expression either across all single cells combined (average log2(TPM)>4.5) or within a single tumor (average log2(TPM)>6 in at least one tumor) were included. Cells expressing less than 2,000 of these 5,948 genes were excluded.

To prove the unstable biopsy, we first divide single-cell RNA-seq data into two groups and performed Kolmogorov-Smirnov test (KS test) [10]. This test distinguishes two different biopsies originated in the same tumor. We performed KS test for 1000 times with random group at each time (Fig 2.a).

The second experiment is to predict a variance of genes' expression (not only EGFR gene) from one single-cell RNA-seq (Fig 2.b). The lesson from EGFR-targeted drug effectiveness implies not average but variance of EGFR expression is important for the resistance of EGFR-targeted drug. We hypothesized the variance of gene X comes from genes related to gene X because a cancer cell secrets and receive signals each other. Thus our hypothesis related genes of a gene X could explain the variance of gene X expression. In this context, gene X is one of 30 genes that are statistically significant compared to normal tissues. We used edge-R to calculate fold-change of genes' expression between a tumor and the normal. We took from top to 30th genes by order of absolute fold-change. Thus we tried to predict the variance from the 29 genes related to a gene X. We used linear least squares and considered to predict 30 genes of heterogeneity. Here we define the heterogeneity of a gene X by the variance of a gene X expression in tumor bulk since definition of clonal heterogeneity is not sophisticated enough for handling them quantitatively. In summary of our date set, input dimension is 29 genes' expression and output dimension is 1 gene's heterogeneity among tumor bulk. Our dataset is 430 single-cell sequencing.

3. Result

In the first experiment, the result of KS test is p-value 0.64±0.02. This indicates it is hard to know whether two groups of single-cell RNA-seq are from the same EGFR gene expression distribution.

In second experiment, we made linear regression models of 30 genes to predict the variance of gene expression. As a result, we obtained 3 genes (IRS2, MAP2, ERBB2) that can predict the heterogeneity with F-test (3 of them have 0.76 score of R-squared). The other 27 genes had low value of R-squared. We analyzed these 3 genes using GeneMania [12]. Using the following graph concept, we found the 3 genes locating in the center of 30 genes (Fig 3). One of possible reasons why those 3 genes are predictable is that they are connected with a good number of different genes and the knowledge on the rest of the other genes is useful enough to identify gene expression

profiles of those 3 genes without any initial knowledge about those 3. This suggests the variance of genes could be predictable from the other genes if genes are located in the center.

Fig 3. Result from GeneMania[12]. Red line above shows physical interactions of genes. 2 genes are connected to 3 genes.

4. Conclusion

This paper investigated the unstableness of biopsy. Traditional procedure of biopsy is not a reliable way to decide EGFR positive or negative. It is extremely useful if we can predict the variance of gene expression from one biopsy. We showed 3 linear regression models which account for the heterogeneity of a gene X and the rest do not work well. Further research is desired to predict other genes.

References

[1] Marusyk A, Almendro V, Polyak K.: Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer.* 2012;12(5):323–34. 10.1038/nrc3261
[2] Hirsch FR, Ottesen G, Pødenphant J, et al. : Tumor heterogeneity in lung cancer based on light microscopic features. A retrospective study of a consecutive series of 200 patients, treated surgically. *Virchows Arch A Pathol Anat Histopathol.* 1983;402(2):147–53. 10.1007/BF00695056

[3] True, L. et al. A molecular correlate to the Gleason grading system for prostate adenocarcinoma. Proc. Natl. Acad. Sci. USA 103, 10991–10996 (2006).
[4] Gleason, D.F. & Mellinger, G.T. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J. Urol. 111, 58–64 (1974).

[5] Wiener RS, Schwartz LM, Woloshin S, Welch HG. Population-based risk for complications after transthoracic needle lung biopsy of a pulmonary nodule: an analysis of discharge records. Ann Intern Med. 2011;155(3):137-144

[6] Ladanyi M, Pao W. Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond. Mod Pathol 2008;21:S16–S22.
[7] Sequist LV, Lynch TJ. EGFR tyrosine kinase inhibitors in lung cancer: an evolving story. Annu Rev Med 2008;59:429–442.

[8] A.P. Patel, I. Tirosh, J.J. Trombetta, A.K. Shalek, S.M. Gillespie, H. Wakimoto, D.P. Cahill, B.V. Nahed, W.T. Curry, R.L. Martuza, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma Science (New York, NY), 344 (2014), pp. 1396–1401

[9] N. E. Navin. The first five years of single-cell cancer genomics and beyond. Genome 860 Research, 25:1499–1507, 2015.
[10] MASSEY, F. J, Jr. The Kolmogorov-Smirnov test of goodness of fit. J. Amer. Statist. (1951), Ass, 46, 68-78

[11] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q Nucleic Acids Res. 2010 Jul 1;38 Suppl:W214-20

[12] Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, et al. . The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function, Nucleic Acids Res. , 2010, vol. 38 Suppl(pg. W214-W220)